DOCUMENT RESUME

ED 433 352 TM 030 021

AUTHOR Aaron, Bruce C.; Kromrey, Jeffrey D.
TITLE Randomization Regression Tests for Single-Subject Data.
PUB DATE 1998-02-00
NOTE 36p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Tampa, FL, February 23-28, 1998).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Simulation; *Effect Size; Monte Carlo Methods; Nonparametric Statistics; *Regression (Statistics)
IDENTIFIERS *Randomization; *Single Subject Designs; Type I Errors; Type II Errors

ABSTRACT

In a Monte Carlo analysis of single-subject data, Type I and Type II error rates were compared for various statistical tests of the significance of treatment effects. Data for 5,000 subjects in each of 6 treatment effect size groups were computer simulated, and 2 types of treatment effects were simulated in the dependent variable during intervention phases, resulting in mean change in level or mean change in slope. Significance test statistics were based on explained variance indicated by squared multiple correlations using multiple regression models that were closely specified to the treatment effect (termed "specific" tests) or that modeled effects beyond those in the data (termed "general" tests). These tests were applied as both parametric and nonparametric (randomization) tests of treatment effects. Results indicate that parametric tests exhibit Type I error control and superior power for independent data, but fail to control Type I error rates for dependent data with autocorrelated observations. In contrast, randomization tests exhibit Type I error control even with serially correlated data, but provide inadequate power for detecting treatment effects and become increasingly conservative with increasing autocorrelation. Implications for analysis of single-subject data series are discussed. (Contains 4 figures, 5 tables, and 38 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 433 352

Randomization regression tests for single-subject data

Bruce C. Aaron

Jeffrey D. Kromrey

University of South Florida

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Bruce Aaron
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM030021

A paper presented at the annual meeting of the
Eastern Educational Research Association, Tampa, Florida, February 1998

BEST COPY AVAILABLE

Abstract

In a Monte Carlo analysis of single-subject data, Type I and Type II error rates were compared for various statistical tests of the significance of treatment effects. Data for 5000 subjects in each of six treatment effect size groups (0, .2, .5, .8, 1.1, and 1.4) were computer-simulated as single-subject time series of 40 observations in an ABAB design with first-order autocorrelations of 0, .10, .20, .30, .50, and .70. Two types of treatment effects were simulated in the dependent variable during intervention phases, resulting in mean change in level or mean change in slope. Significance test statistics were based on explained variance indicated by squared multiple correlations (R^2), using multiple regression models that either were closely specified to the treatment effect (termed specific tests) or that modeled effects beyond those in the data (termed general tests). These tests were applied as both parametric (conventional F-tests) and nonparametric (randomization) tests of treatment effects. Results indicate that parametric tests exhibit Type I error control and superior power for independent data, but fail to control Type I error rates for dependent data with autocorrelated observations. In contrast, randomization tests exhibited Type I error control even with serially correlated data, but provided inadequate power for detecting treatment effects, and became increasingly conservative with increasing autocorrelation. In addition, little difference was found between the performance of specific randomization tests and general randomization tests. The results suggest that researchers concerned with the statistical analysis of similar single-subject data series face the dilemma of (a) using randomization test procedures which conservatively control Type I error, regardless of autocorrelation, but provide inadequate levels of statistical power, or (b) using traditional parametric procedures which provide adequate power but fail to control Type I error rates when data are autocorrelated. Alternative strategies might include foregoing hypothesis testing of such single-subject data series, and using instead a sample estimate of the effect size.

Randomization regression tests for single-subject data

Background

The application of scientific method to distinctly human affairs, as embodied in the social sciences, presumes that human behavior can be summarized usefully by functional relationships among measurable variables. The covenant of the disciplines of psychology and education for improving the human state of affairs rests upon the elucidation and generalizability of these systematic relationships. Some, however, have questioned the efficacy of behavioral science, including the field of educational research, to discover such durable, generalizable relationships (Gage, 1996). Kessels and Korthagen (1996), for example, note the notoriously ineffective translation of educational theory into practice, and attribute it to conflicting perspectives of rationality that have been apparent since the beginning of Western philosophy. These perspectives are represented by Plato's conception of knowledge as episteme (i.e., propositional, generalizable, conceptual, and abstract) on the one hand, and Aristotle's conception of knowledge as phronesis (i.e., practical, situational, perceptual, and concerned with the particular concrete case), on the other. The latter directly emphasizes the individual, idiosyncratic single case, while epistemic knowledge is held to address problems of the single case only to the extent that the individual is an exemplar of the more general type or category (Donmoyer, 1996). While phronesis and episteme are not mutually exclusive, Kessels and Korthagen (1996) suggest that a relative emphasis on the latter prevents a potent link between the collection of knowledge through educational research and the application of knowledge in educational practice. The sense of dissonance between the search for knowledge of the singular and knowledge of the general, exemplified in these Aristotelian and Platonic conceptions of knowledge, seems reflected in separate modern paradigms of single-subject and group comparison research in psychology and education.

The scientific study of human behavior began with data gathered from repeated measures taken on individuals over time. Near the beginning of the twentieth century, however, the foundation had been laid for the application of statistical procedures to the psychological and educational measurement of individual differences, which ultimately supported the current framework of group comparison approaches used in modern behavioral research.

Dependence on group comparison research methodology, however, has drawn detractors during the latter half of the twentieth century, particularly among researchers in applied behavior analysis who have been proponents of single-subject repeated measures designs (Skinner, 1953, Sidman, 1960). For applied behavior analysis, the establishment of functional relations between

independent and dependent variables traditionally relies on repeated measures and within-subject experimental designs (Poling & Grossett, 1986).

When based on the analysis of the single case, the study of human behavior is referred to as idiographic (from the Greek, *idios*, "one's own; private"), while analyses based on group data are termed nomothetic (from the Greek, *nomos*, "law"). According to proponents of idiographic methods, limitations inherent in the nomothetic approach derive from ethical considerations, logistic obstacles, the obfuscation of clinically significant individual outcomes, difficulties in generalization to individuals, and a neglect of within-subject variability (Barlow & Hersen, 1984). Similarly, advocates of single-subject research emphasize that: a) the processes of cognition and behavior are continuous, b) these processes occur at the individual level, and therefore should be investigated at the level of the individual, and c) between-group analyses of summary measures based on means and standard deviations provide limited information about the effects of independent variables at the level of individual functioning (Kratochwill, 1992). Concerns about inadequacies of nomothetic research and the benefits of idiographic approaches are apparent in an increased recent interest in single-subject research as reflected in the professional literature, noted by Barlow and Hersen (1984) and Kratochwill (1992).

Design and analysis of single-subject data

Four major designs are common in single-subject research: the AB, the withdrawal or reversal (e.g., ABA, ABAB), multiple baseline, and alternating treatments. Onghena (1992) noted that the ABA (or similarly, the ABAB) withdrawal or reversal design is the most typical of repeated measures single-subject designs. In the ABAB design, the initial phase, in which the dependent variable is measured under conditions that do not include the independent variables of interest, provides a baseline (A). In the second phase, the independent variable is introduced (B). In the third phase, threats to internal validity are tested by withdrawal of the independent variable (A). Finally, the independent variable is re-introduced (phase B). A change in behavior during the B intervention phases provides evidence of a functional relationship between the independent and dependent variables. Variations of this withdrawal, or reversal, design (e.g., ABA, ABACAB, ABABAB) have similar characteristics and interpretation.

Regardless of the design used, data are typically submitted to either visual analysis or statistical analysis. Of course, statistical analysis does not preclude visual representation and analysis of data, although many researchers claim that visual analysis alone is appropriate. Each approach has proponents and detractors among researchers (Baer, 1977; DeProspero & Cohen, 1979; Gottman & Glass, 1978; Michael, 1974; Skinner, 1966; Wampold & Furlong, 1981), and

problems attributed to each approach have led to considerable discussion regarding the most effective analysis method.

Visual analysis

Visual analysis is the primary method used in studies published in the leading journals of the field, (e.g. The Journal of Applied Behavior Analysis) (Kratochwill, 1992). It provides a direct, concise, and complete representation of the nature of the experiment and the results of the research by a single graphic display that indicates the sequence and length of conditions and the sequential values of the dependent measures.

Parsonson and Baer (1992) cite advantages, relative to statistical analysis, of graphing single-subject data. Among their arguments are that visual analysis is quick, convenient, flexible, and accessible to students and researchers of varying levels of expertise. In addition, they maintain that graphic analysis provides the least transformative representation of the data as actually measured, and lacks the uncertainty and complexity inherent in statistical analyses.

Visual analysis, however, is subject to disagreement across analysts about the effects of interventions, including disagreement about whether reliable effects have occurred and about interpretation of the data (DeProspero & Cohen, 1979; Gottman & Glass, 1978; Jones, Vaught, & Weinrott, 1978; Ottenbacher, 1986; Sharpley, 1981).

Statistical analysis

Statistical analysis of single-subject data circumvents some problems that plague visual analysis, in that statistical methods will produce consistent results across data analysts and provide a standard or criterion on which to base the magnitude of treatment effects. However, when different statistical methods are applied to the same data, the agreement rates across methods are not exceptionally high (Nourbakhsh & Ottenbacher, 1994). In addition, it is often difficult to determine an appropriate statistical analysis to use with unstable baseline data and with serially dependent (autocorrelated) data.

Traditional statistical analyses, such as analysis of variance (ANOVA), have been recommended as models for determining whether or not a single case time series exhibits change (Chassan, 1967; Borg & Gall, 1989). ANOVA, however, assumes that observations (specifically, error components of observations) are independent. But since each observation in a time series (except the first) might be dependent to some extent on the value of the observation preceding it, Type I error rates rise to the extent that this autocorrelation is not controlled (Ostrom, 1990). Hence, conventional F and t tests have been recommended for use only when autocorrelation is

found to be insignificant (Kazdin, 1984). However, the prevalence of significant autocorrelation in behavioral data is itself disputed. Huitema and McKean (1994a, 1994b), for example, claim that significant autocorrelation has been severely overestimated in behavioral time-series data, and that conventional tests of the first-order autocorrelation are biased for small samples (which are common in applied behavioral analysis studies).

Randomization testing. One interesting statistical approach to the analysis of single-subject experiments is provided by randomization testing, a nonparametric technique for testing the statistical significance of treatment effects which assumes random assignment within the design (such as the random assignment of treatments to times) (Edgington, 1992). Conducting significance tests with randomization test procedures provides a number of advantages in testing hypotheses about functional relations of variables in experiments, including weak assumptions about the distribution of population parameters of interest.

Random assignment is the primary design strategy employed to meet the fundamental need of researchers in psychology and education to make the case that conditions manipulated in experiments are responsible for results that tend to refute a null hypothesis. In research with groups, this entails random assignment of subjects to treatment groups. In single-subject experiments, equivalent design strategies are the random assignment of treatment group to occasions/times, or the random assignment of the onset of treatments. The null hypothesis for randomization tests of treatment effects for data of a single subject across time is that of no difference in effect for the randomly assigned experimental units (i.e., treatment times).

Hypothetically, if the behavior generating the observed series data values is unaffected by the random assignment of treatments to times, then the random selection of other occasions as the beginning and ending points of treatments would not affect the observed data series for the subject. In a randomization test, the test statistic (e.g., the differences in means for treatment phases within the series), which can be tailored to meet the given situation, is chosen and then calculated for the observed data. The data are permuted, based on the other equally-probable random assignments that could have been selected given the randomization scheme, and the test statistic derived for each of these permutations generates a distribution (e.g., of t or F). The proportion of test statistics as large or larger than the observed statistic provides the significance level. Essentially, then, the steps for a single-subject randomization test are as follows:

1) Choose a statistic to allow comparison of a null and alternative hypothesis.
2) Compute the statistic for the observed data sample.

3) Permute the treatment phases in another equally probable random way that might have resulted from the randomization scheme chosen for the experiment, and compute the statistic for the permutation of the data values. The raw data values themselves are not permuted and their order is thus preserved.
4) Repeat step 3 until all permutations (or an acceptable subset of them) have been derived and a test statistic has been calculated for each rearrangement of the data.
5) Count the number of test statistics in the permutation distribution generated by step 4 that are equal to, or more extreme than, the statistic for the observed data. This number, divided by the number of permutations conducted, provides the significance level of the test.

Since randomization tests rely on an empirically derived permutation distribution, rather than a theoretical distribution, they are recommended as statistical tests for situations in which assumptions of normality, homogeneity of variance, and random sampling, inherent in parametric significance tests, cannot be justified. Based on observed data, the randomization test is distribution-free like other nonparametric tests that rely on rank transformations of observed scores (e.g., the Mann-Whitney U or the Kruskal-Wallis), but preserves the scale values of the data.

In general, the more specific the alternative hypothesis and statistical test the more powerful the randomization test (Onghena, 1992). Often therefore, specific test statistics are recommended for their likely sensitivity to the anticipated treatment effect. Accurate prediction of the treatment effect, however, might not be a practical requirement for all single-subject research. In addition, it is unknown at this time to what extent specific tests do indeed improve Type I error control, or increase power. Ferron (1993) suggested that a more general test could be based on the sum of the proportions of variance explained (R^2) by regressions within each intervention phase. Similarly, Kromrey and Foster-Johnson (1996) demonstrated the utility of R^2 in calculating effect sizes as descriptive statistics for single-subject data. The current study adopted these models and investigated the efficacy of using magnitudes of R^2 as a test statistic in an inferential approach to the analysis of single-subject data. Explained variance (R^2) was used to provide a specific statistical test of treatment effects for cases in which the regression model adopted was specifically compatible with the data, and served as a general test statistic for cases in which the regression model specified included terms that were be sensitive to a wider range of effects than those apparent in the observed data. Evidence suggests that when applying these tests as randomization tests for single-subject experiments, a more general, flexible test can perform similarly to a specific test for treatment effects (Aaron, Kromrey, & Foster-Johnson,

1996). However, additional research was needed to help determine whether this functionality holds across varying levels of autocorrelation, effect size, and type of treatment effect.

Type I error rates can be reasonably controlled by randomization tests, even in the presence of autocorrelation, but a lack of power for these tests has been demonstrated under certain conditions (Edgington, 1980; Ferron, 1993; Ferron & Ware, 1994; Aaron, Kromrey, & Foster-Johnson, 1996). Since randomization tests might provide a feasible alternative to conventional statistical analysis, stochastic modeling, or reliance on visual means of assessing the effects of interventions on individuals, it is important to determine the conditions under which randomization testing is effective in single-subject analyses. Specifically, this study examined the power and Type I error control of general and specific test statistics, used within both parametric and randomization tests, across varying treatment effect sizes, for both autocorrelated and independent single-subject observations in a common single-subject design (ABAB).

Method

Design and Analysis

The success of general versus specific test statistics and the effectiveness of randomization tests relative to parametric tests of single-subject data were evaluated in a Monte Carlo study. Five thousand single-subject samples were simulated for each of six effect sizes: 0 (the null model), .2, .5, .8, 1.1, and 1.4. These levels represent treatment effects ranging from small to large as suggested by Cohen (1988), and were extended to encompass larger treatment effect sizes that are common in published results of single-subject studies (Foster-Johnson, 1997). These treatment effect sizes were crossed with six magnitudes of autocorrelation (0, .1, .2, .3, .5, and .7). In the null effect size condition, no population mean differences were introduced across the ABAB phases in the 40-observation series. Two types of treatments effects were simulated to occur within the five non-null conditions. The first was change in level, with a constant added to each observation in the second and fourth phases (the B conditions) in the ABAB series. The magnitude of the constant was varied to produce the desired effect size. The second type of effect was change in slope, generated by multiplying observations in intervention phases by a vector of monotonically increasing values to produce each of the five non-null effect sizes as a change in slope, with a constant mean level.

The study is represented as a 2 x 2 x 2 x 6 x 6 factorial design, with respective levels of type of test (parametric v. randomization), specificity of regression model for measuring explained variance (specific v. general), type of treatment effect (change in level v. change in slope), magnitude of treatment effect (effect sizes of 0, .2, .5, .8, 1.1, and 1.4) and magnitude of autocorrelation (0, .1, .2, .3, .5, and .7) chosen as levels of the independent variables. The core design, depicted in Table 1, specifies 144 experimental conditions, examined for two different data sets. The first data set demonstrated change-in-level effects, and the second demonstrated change-in-slope effects. This resulted in 288 experimental conditions. Each condition was applied to data for 5000 simulated subjects. Each single-subject sample was comprised of 40 serial observations (10 each in an ABAB design). Serial correlations of .1, .2, .3, .5, and .7 were constructed as first-order autoregressive, ARIMA (1,0,0).

Insert Table 1 about here

Model Specification. The general and specific statistical tests indicated in Table 1 were derived using regression models to determine the significance probability of effects for each data series. For the parametric tests, change in R^2 was derived through hierarchical multiple regression, a multiple-step procedure. For each case, the relationship of Y to the lag, T (the sequential observation points), accounts for trend across the series, if trend is present. This is represented by b_1 in the equation:

$$Y = b_0 + b_1 T + e \qquad (1.)$$

where b_0 represents the initial value of Y and T equals the lag (T =0 for the first observation within a series). This model ignores treatment phases as an independent variable. Subsequently, a regression equation is derived that will fit the data while taking into account the treatment phases of the ABAB design. These phases can be described as values in a dummy vector, X_1 (where X_1 takes the value of 0 for phase A_1 and A_2 , and 1 for B_1 and B_2) as shown below:

$$Y = b_0 + b_1 T + b_2 X_1 + e \qquad (2.)$$

Since both the trend across all the data ($b_1 T$) and variance explained by between-phase differences ($b_2 X_1$) are modeled, equation 2 models the change in level between phases after adjusting for

trend. Essentially, this allows separate regression equations to be fitted for each treatment condition, with a common *b*, but different intercepts. An R^2 is derived for each of the two equations above (R^2_1 and R^2_2, respectively). The incremental variance accounted for by employing the full model is tested for statistically significant difference from zero by the following equation:

$$F = \frac{(R^2_2 - R^2_1) / (k_2 - k_1)}{(1 - R^2_2) \quad / \quad (N - k_2 - 1)} \qquad (3.)$$

This test would be specifically sensitive to changes in level for the five non-zero effect sizes (.2, .5, .8, 1.1, and 1.4), and thus provided the specific parametric test for the data that demonstrate a treatment effect of change in level. The general parametric test for these data was conducted testing incremental variance accounted for by creating a full model with the addition of a multiplicative interaction term, as in:

$$Y = a + b_1T + b_2X_1 + b_3TX_1 + e \qquad (4.)$$

The inclusion of the product vector, b_3TX_1, accounts for the interaction between treatment and time, modeling differences in slope for different treatment conditions, as well as differences in level. The use of this multiplicative linear regression model provides a general test for effects exhibiting a mean change in level effect, since it tests for change in both level and slope while adjusting for trend. Here, the model represents a differential effect of the interaction or moderating variables (e.g., X_1 as a dummy-coded variable representing an A or B treatment condition), resulting in different slopes of T on Y for different values of the moderator variable, and allowing the slopes within treatment phases of the time-series to vary. This interaction effect is not included in the restricted additive model (equation 2), which fits a common slope for the intervention phases.

For treatment effects that result in a change in slope in the data, incremental variance accounted for by the addition of the multiplicative interaction term in equation 4 provides the specific test, since it fits, or explains, variance due to slopes that vary across intervention phases, reflecting the effect of treatment. This was the specific test, therefore, for the parametric tests used to examine data exhibiting changing slopes during treatment. By extension, the general parametric test for the change-in-slope treatment effect conditions is provided by the significance of incremental variance accounted for by the addition of the quadratic polynomial term, as in:

$$Y = a + b_1T + b_2X_1 + b_3TX_1 + b_4T^2X_1 + e, \qquad (5.)$$

since it allows representation of a more complex, curvilinear form of the interaction. The model summarized by equation 5 expresses a curvilinear moderated relationship between T and X_1, which is sensitive not only to slopes that vary across A and B treatment phases, but to a moderated relationship between X_1 and T that exhibits curvilinearity. A quadratic relationship (involving a curve with a single bend), expressing a changing rate of increase or decrease in the trend of the data, would be represented in the full regression model of equation 5.

Specific and general regression models were applied in both the parametric and randomization testing conditions (see Table 1). The parametric models determined statistical significance by reference to the theoretical distribution of *F*, and used incremental explained variances as test statistics, as described above. For each nonparametric randomization test, on the other hand, an empirical randomization distribution was built for each of the 5000 samples, and the R^2 test statistic was employed without reference to incremental explained variance derived by comparison to a restricted regression model. Thus, for the randomization tests used in the study, equation 2 was used to derive the specific test for data showing treatment effects as mean level changes, and equation 4 provided the general test of that data. For treatment effects resulting in changes in slope, the R^2 test statistic produced by equation 4 provided the specific test, and equation 5 provided the general test of the data. Procedurally, these randomization tests determined significance by comparing the particular model's explanation of variance in the observed data series to the R^2 produced by fitting the model to each of the possible permutations of the data in the series. The randomization distribution for each randomization test contained 165 permuted R^2 statistics as a result of the randomization scheme employed, which was set at a minimum of eight observations per phase. Given the constraints of four treatment phases and 40 observations, this resulted in 165 possible permutations, and a minimum achievable significance probability of .006. Of particular interest was whether the use of general tests (defined by regression models that are sensitive to a greater array of treatment effects) sacrifices power or sensitivity in detecting treatment effects for single-subject experiments.

All program code was written in SAS version 6.11. The data generation algorithm used SAS/Interactive Matrix Language (SAS/IML), based on modifications to the ARMA (Autoregressive Moving Average) program included in the SAS/IML User's Guide (p. 150-151). The data generation algorithm was verified by calculating the jackknife estimate of sample autocorrelation (Huitema & McKean, 1994a). The rest of the code was verified by checking the calculations with a benchmark set of data.

Results

For each experimental condition, the proportion of tests (5000 per cell) that indicated statistically significant changes for treatment effect was tabulated. For each case in the parametric test experimental conditions, significance probability was determined by referring to the point in the standard distribution of *F*-ratios at which the observed test statistic fell. For each case in the randomization test experimental condition, significance probability was determined by referring to the point in the empirically derived randomization distribution at which the observed test statistic fell. The results across data sets demonstrating a mean shift effect for treatment are displayed in Tables 2 and 3, were calculated based on nominal alpha levels of .05 and .10.

Insert Tables 2 and 3 about here

The results for treatment effects demonstrating a change in slope are displayed in Tables 4 and 5, calculated based on nominal alpha levels of .05 and .10.

Insert Tables 4 and 5 about here

As anticipated, the parametric tests did not adequately control Type I error rates with serially dependent data. The relationships of autocorrelation to Type I error for each type of test are depicted in Figure 1, for mean-shift tests. The relationships of autocorrelation to Type I error are similar for each type of test for treatment effects resulting in slope change .

Insert Figure 1 about here

The power levels of those experimental conditions in which Type I error was not inflated were examined across the varying treatment effect sizes. The conditions controlling Type I error were the parametric tests with independent data (i.e., with no autocorrelation), and the randomization tests. For these, power is indicated by the proportions of the 5000 tests per cell in

which the null hypothesis was rejected. The levels of power found across effect sizes for independent data is displayed for each specific and general test of mean-shift data in Figure 2, and for slope-changing data in Figure 3.

Insert Figures 2 and 3 about here

The randomization tests controlled Type I error but were increasingly conservative with increasing autocorrelation, and exhibited significantly less power than the parametric tests for data with no autocorrelation. This conservatism is shown in Figure 4, which displays the effect of autocorrelation on null hypothesis rejection rates at each treatment effect size for specific randomization tests of mean-shift treatment effects.

Insert Figure 4 about here

Only a small difference in performance was indicated between the general and specific tests across the six effect sizes for both the null and autocorrelated models, whether conducted with parametric or randomization tests. Interestingly, the specific tests provided little additional power. For the randomization tests, power to detect treatment effects was not practically adequate.

According to the parameters of the simulated data sets, treatment effects were modeled as changes in level and changes in slope, respectively. With only random differences in treatment slopes for the mean-shift data, or in the linearity of changing slopes for the slope-change data, little difference in regression sums of squares resulted between conditions when employing the additional explanatory parameters of the fuller general models in the explanation of dependent variable scores. The statistical tests of whether greater explanation of variance in scores results from using the additional vectors of the full models were generally insignificant. The difference in the specific and general regression tests, when the population from which samples are drawn consists of equal slopes and different mean treatment levels for the mean-shift data, or differing slopes for the slope-change data, consists of the loss of one degree of freedom from the

denominator of the F statistic. For these data, this loss did not result in a substantial loss of statistical power.

Discussion

The results of this research are congruent with previous studies that have pointed out the power deficits of randomization tests applied to single-subject data (Ferron, 1993, Ferron & Ware, 1994). Further, the lack of Type I error control evidenced by the parametric procedures is in agreement with previous studies (Greenwood & Matyas, 1990; Toothaker, Banz, Noble, Camp & Davis, 1983).

The results confirm previous evidence (Aaron, Kromrey, & Foster-Johnson, 1996) suggesting that when using these hierarchical randomization tests to analyze data from single-subject experiments, a more general, flexible test can perform similarly to a specific test for treatment effects. The current results, however, strongly suggest that this flexibility is a moot advantage given the unacceptably low power provided by the randomization tests (at least for the conditions reflected in these data).

The current results suggest greater sensitivity and power across the experimental conditions for tests of changing slope treatment effects relative to mean shift treatment effects. However, the effect sizes for these qualitatively different treatment effects are calculated differently, and were generated for each of the simulated data sets according to the formulas prescribed for each effect. Conclusions regarding differences in the power between tests of the mean-shift data and slope-change data sets should be deferred until the equivalence in scale for these different effects sizes can be confirmed through further investigation.

The results suggest that researchers concerned with the statistical analysis of similar single-subject data series face the dilemma of (a) using randomization test procedures which conservatively control Type I error, regardless of autocorrelation, but provide inadequate levels of statistical power, or (b) using traditional parametric procedures which provide adequate power but fail to control Type I error rates when data are autocorrelated. One may seek solace in the argument of Huitema (1985) that autocorrelation is rare in single-subject data, but apparently few researchers have done so (Ferron & Ware, 1994). An alternative strategy would be to avoid entirely the testing of null hypotheses and to use instead a sample estimate of the effect size, a procedure recommended by Kromrey and Foster-Johnson (1996).

Of interest was the question of how large autocorrelation can be before the traditional parametric tests lose control of Type I error rates. This study suggests that even at a low level of autocorrelation (.1), traditional parametric procedures cannot be presumed to control Type I error. A related area of inquiry is the identification of autocorrelation in samples. The jackknife procedure detailed by Huitema and McKean (1994a) provides an unbiased estimate, but the power of this procedure has not been established.

In addition, this study examined two particular types of treatment effect (i.e., a shift effect, or change in level, and a trend effect, or change in slope). Additional types of treatment effects such as changes in variability represent viable alternative hypotheses in single-subject research. Examining the performance of these types of tests with such treatment effects might be worthwhile.

An increased interest among educational and psychological researchers in single-subject studies has enlarged and refined the set of research methods and analysis techniques applicable to the single case study. Applications of single-subject analyses might also be fostered by the growing demand for alternative, frequent assessment techniques, borne of the perceived inadequacies of annual standardized norm-referenced tests (FCERA, 1994). As educators seek to expand their methods of student assessment, single-subject analyses might assume a more prominent and useful position among methods of educational measurement and research. It is hoped that this study can help inform such research efforts by contributing to our knowledge of the effectiveness of particular tools for the analysis of single-subject data, particularly concerning the limitations of randomization testing and traditional parametric testing of treatment effects.

References

Aaron, B.C., Kromrey, J.D., & Foster-Johnson, V.L. (1996). Parametric and nonparametric analyses of single-subject data. Paper presented at the nineteenth annual meeting of the Eastern Educational Research Association, Cambridge, Massachusetts.

Baer, D. (1977). Perhaps it would be better not to know everything. Journal of Applied Behavior Analysis, 10, 167-172.

Barlow, D.H. & Hersen, M. (1984). Single case experimental designs. Strategies for studying behavior change. New York: Pergamon Press.

Borg, W.R., and Gall, M.D. (1989). Educational research: An introduction. New York: Longman.

Chassan, J.B. (1967). Research design in clinical psychology and psychiatry. New York: Irvington.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: LEA.

DeProspero, A., and Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. Journal of Applied Behavior Analysis, 12, 573-579.

Donmoyer, R. (1996). Juxtaposing articles / posing questions: An introduction and an invitation. Educational Researcher, 5, (3), 4.

Edgington, E.S. (1980). Validity of randomization tests for one-subject experiments. Journal of Educational Statistics, 5, 261-267.

Edgington, E.S. (1992). Nonparametric tests for single-case experiments. In T.R. Kratochwill & J.R. Levin (Eds.) Single-Case Research Design and Analysis. LEA Associates: Hillsdale, New Jersey.

Ferron, J. (1993). Suggested solutions to problems facing the use of randomization tests with single-case designs. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Ferron, J, and Ware, W. (1994). Analyzing single-case data: How powerful are randomization tests? Paper presented at the annual meeting of the American Educational Research Association.

Florida Commission on Education Reform and Accountability (1994). *Blueprint 2000: A system of school improvement and accountability.*

Foster-Johnson, L. (1997). Apples and oranges? Comparative interpretation of effect sizes in single subject research. Paper presented at the annual meeting of the Eastern Educational Research Association.

Gage, N.L. (1996). Confronting counsels of despair for the behavioral sciences. Educational Researcher, 5, (3), 5-15.

Gottman, J.M. & Glass, G.V. (1978). Analysis of interrupted time-series experiments. In T.R. Kratochwill (Ed.), Single-subject research. Strategies for evaluating change (pp. 197-235). New York: Academic Press.

Huitema, B.E. (1985). Autocorrelation in applied behavior analysis: A myth. Behavioral Assessment, 7, 107-118.

Huitema, B.E., and McKean, J.W. (1994a). Reduced bias autocorrelation estimation: Three jackknife methods. Educational and Psychological Measurement, 54, 654-665.

Huitema, B.E., and McKean, J.W. (1994b). Tests of H_o: $p_1 = 0$ for autocorrelation estimators r_{F1} and r_{F2}. Perceptual and Motor Skills, 78, 331-336.

Jones, R.R., Weinrott, M.R., & Vaught, R.S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. Journal of Applied Behavioral Analysis, 11, 277-283.

Kazdin, A.E. (1984). Statistical analyses for single-case experimental designs. In D.H. Barlow & M. Hersen (Eds.) Single-case experimental designs: Strategies for studying behavior change New York: Pergamon Press.

Kessels, J.P.A.M., & Korthagen, F.A.J. (1996). The relationship between theory and practice: Back to the classics. Educational Researcher, 5, (3), 17-22.

Kratochwill, T.R. (1992). Single-case research design and analysis: An overview. In T.R. Kratochwill and J.R. Levin (Eds.), Single-Case Research Design and Analysis. Hillsdale, New Jersey: LEA.

Kromrey, J.D. and Foster-Johnson, V.L. (1996). Determining the efficacy of intervention: The use of effect sizes to augment interpretation of single subject research. Journal of Experimental Education.

Matyas, T.A. & Greenwood, K.M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. Journal of Applied Behavior Analysis, 23, 341-351.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? Journal of Applied Behavior Analysis, 7, 647-653.

Nourbakhsh, M.R. and Ottenbacher, K.J. (1994). The statistical analysis of single-subject data: A comparative examination. Physical Therapy, 74, 768-776.

Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: A rejoinder. Behavioral Assessment, 14, 153-171.

Ostrom, C.W. (1990). Time series analysis: regression techniques. (2nd ed) Sage university paper series on quantitative applications in the social sciences, 07-009. Newbury Park, CA: Sage.

Ottenbacher, K. (1986). Reliability and accuracy of visually analyzing graphed data from single-subject design. American Journal of Occupational Therapy, 40, 464-469.

Parsonson, B.S., & Baer, D.M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T.R. Kratochwill & J.R. Levin (Eds.) Single-Case Research Design and Analysis. LEA Associates: Hillsdale, New Jersey.

Poling, A., and Grossett, D. (1986). Basic research designs in applied behavior analysis. In Poling and R.W. Fuqua (Eds.) Research methods in applied behavior analysis (pp. 7-27). New York: Plenum Press.

SAS Institute, Inc. (1988). SAS/IML software: Usage and reference, version 6. Cary, NC: SAS Institute.

Sharpley, C. (1981). Time series analysis of counseling research. Measurement and Evaluation in Guidance, 14, 149-157.

Skinner, B.F. (1953). Science and Human Behavior. New York: Macmillan.

Skinner, B.F. (1966). Operant behavior. In W.K. Honig (Ed.), Operant Behavior: Areas of research and application. New York: Appleton-Century-Crofts.

Sidman, M. (1960). Tactics of Scientific Research: Evaluating Experimental Data in Psychology. New York: Basic Books.

Toothaker, L.E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). The reliability and accuracy of time series model identification. Educational Review, 7, 551-560.

Wampold, B.E and Furlong, M.J. (1981). The heuristics of visual inference. Behavioral Assessment, 3, 93-103.

Table 1. Study Design
(unshaded cells = changing level data; shaded cells = changing slope data; n = 5000)

Focus of Tests	*Effect Size*	***Parametric Tests***						***Randomization Tests***					
		autocorrelation											
		.00	.10	.20	.30	.50	.70	.00	.10	.20	.30	.50	.70
Specific	.00												
	.20												
	.50												
	.80												
	1.1												
	1.4												
General	.00												
	.20												
	.50												
	.80												
	1.1												
	1.4												

BEST COPY AVAILABLE

Table 2 Null hypothesis rejection rates. Nominal alpha = .05; Effect type = Mean Shift. n=5000.

Focus of Tests	*Effect Size*	***Parametric Tests***						***Randomization Tests***					
		autocorrelation											
		.00	*.10*	*.20*	*.30*	*.50*	*.70*	*.00*	*.10*	*.20*	*.30*	*.50*	*.70*
Specific	*.00*	.054	.078	.116	.138	.234	.351	.031	.032	.039	.023	.016	.016
	.20	.072	.105	.136	.173	.258	.360	.050	.041	.034	.032	.036	.021
	.50	.210	.240	.277	.288	.369	.414	.109	.111	.076	.090	.068	.038
	.80	.493	.490	.509	.503	.521	.510	.227	.194	.177	.158	.127	.086
	1.1	.765	.774	.750	.732	.691	.631	.350	.341	.292	.280	.213	.143
	1.4	.945	.926	.905	.892	.823	.736	.549	.517	.465	.433	.345	.242
General	*.00*	.053	.085	.132	.186	.339	.516	.030	.032	.031	.022	.016	.017
	.20	.064	.102	.153	.207	.358	.520	.042	.034	.035	.030	.032	.030
	.50	.178	.218	.273	.317	.450	.564	.097	.085	.062	.077	.054	.036
	.80	.419	.442	.476	.513	.575	.631	.210	.157	.167	.137	.108	.073
	1.1	.697	.714	.713	.723	.723	.723	.328	.310	.289	.277	.187	.124
	1.4	.906	.894	.878	.877	.835	.793	.530	.491	.430	.410	.323	.209

Table 3. Null hypothesis rejection rates. Nominal alpha = .10 ; Effect type = Mean Shift. n=5000

Focus of Tests	*Effect Size*	***Parametric Tests***						***Randomization Tests***					
		autocorrelation											
		.00	*.10*	*.20*	*.30*	*.50*	*.70*	*.00*	*.10*	*.20*	*.30*	*.50*	*.70*
Specific	*.00*	.107	.139	.192	.222	.329	.451	.070	.065	.070	.055	.040	.042
	.20	.136	.173	.217	.259	.356	.456	.095	.086	.078	.084	.065	.047
	.50	.318	.349	.381	.404	.472	.503	.192	.190	.149	.150	.112	.101
	.80	.626	.616	.627	.617	.610	.598	.366	.332	.314	.279	.217	.152
	1.1	.858	.857	.832	.814	.765	.708	.516	.517	.463	.443	.354	.243
	1.4	.974	.960	.944	.933	.875	.796	.724	.677	.632	.620	.506	.378
General	*.00*	.106	.152	.212	.285	.447	.612	.068	.062	.062	.054	.033	.045
	.20	.134	.181	.243	.309	.469	.612	.084	.081	.072	.076	.078	.060
	.50	.274	.326	.386	.428	.557	.659	.176	.150	.145	.130	.108	.081
	.80	.549	.570	.603	.630	.674	.712	.346	.292	.284	.247	.193	.128
	1.1	.803	.816	.806	.811	.794	.786	.495	.481	.421	.415	.302	.210
	1.4	.955	.943	.929	.925	.884	.851	.717	.652	.608	.578	.467	.327

Table 4. Null hypothesis rejection rates. Nominal alpha = .05; Effect type = slope change. n=5000

Focus of Tests	*Effect Size*	***Parametric Tests***						***Randomization Tests***					
		autocorrelation											
		.00	*.10*	*.20*	*.30*	*.50*	*.70*	*.00*	*.10*	*.20*	*.30*	*.50*	*.70*
Specific	*.00*	.053	.085	.132	.186	.339	.516	.034	.030	.029	.025	.022	.011
	.20	.153	.196	.235	.299	.438	.583	.084	.074	.063	.063	.040	.027
	.50	.680	.685	.699	.704	.727	.742	.340	.319	.288	.250	.185	.106
	.80	.974	.970	.962	.954	.931	.900	.683	.656	.601	.560	.421	.272
	1.1	1.0	.998	.999	.997	.985	.967	.882	.861	.848	.787	.660	.485
	1.4	1.0	1.0	1.0	1.0	.997	.989	.958	.949	.934	.909	.797	.619
General	*.00*	.050	.095	.146	.207	.369	.546	.028	.025	.022	.020	.017	.010
	.20	.132	.182	.236	.302	.464	.609	.063	.052	.052	.045	.029	.017
	.50	.621	.645	.673	.681	.731	.745	.262	.224	.207	.181	.124	.060
	.80	.959	.957	.951	.945	.923	.896	.542	.522	.465	.427	.308	.180
	1.1	.999	.996	.998	.995	.983	.964	.772	.747	.711	.657	.517	.350
	1.4	1.0	1.0	1.0	1.0	.997	.985	.886	.863	.843	.821	.686	.466

Table 5. Null hypothesis rejection rates. Nominal alpha = .10; Effect type = slope change. n=5000.

Focus of Tests	*Effect Size*	***Parametric Tests***						***Randomization Tests***					
		autocorrelation											
		.00	*.10*	*.20*	*.30*	*.50*	*.70*	*.00*	*.10*	*.20*	*.30*	*.50*	*.70*
Specific	*.00*	.106	.152	.212	.285	.447	.612	.076	.064	.064	.056	.049	.033
	.20	.245	.301	.343	.402	.540	.671	.156	.135	.127	.117	.079	.057
	.50	.787	.792	.799	.791	.802	.816	.493	.469	.430	.392	.293	.182
	.80	.990	.987	.982	.974	.959	.933	.828	.796	.751	.710	.564	.388
	1.1	1.0	1.0	1.0	.999	.991	.983	.954	.936	.916	.891	.806	.618
	1.4	1.0	1.0	1.0	1.0	.999	.994	.988	.986	.978	.964	.891	.739
General	*.00*	.104	.165	.237	.311	.491	.639	.061	.059	.057	.051	.044	.028
	.20	.224	.286	.354	.420	.578	.696	.119	.102	.101	.092	.064	.040
	.50	.737	.761	.768	.779	.807	.820	.388	.356	.323	.288	.209	.112
	.80	.982	.978	.976	.968	.954	.932	.705	.677	.621	.577	.437	.278
	1.1	1.0	.999	1.0	.999	.991	.978	.882	.852	.832	.786	.665	.470
	1.4	1.0	1.0	1.0	1.0	.999	.993	.953	.945	.924	.907	.793	.593

Figure 1.

Type I Error Rates, Randomization v. Parametric Tests
(null effect, alpha = .05)

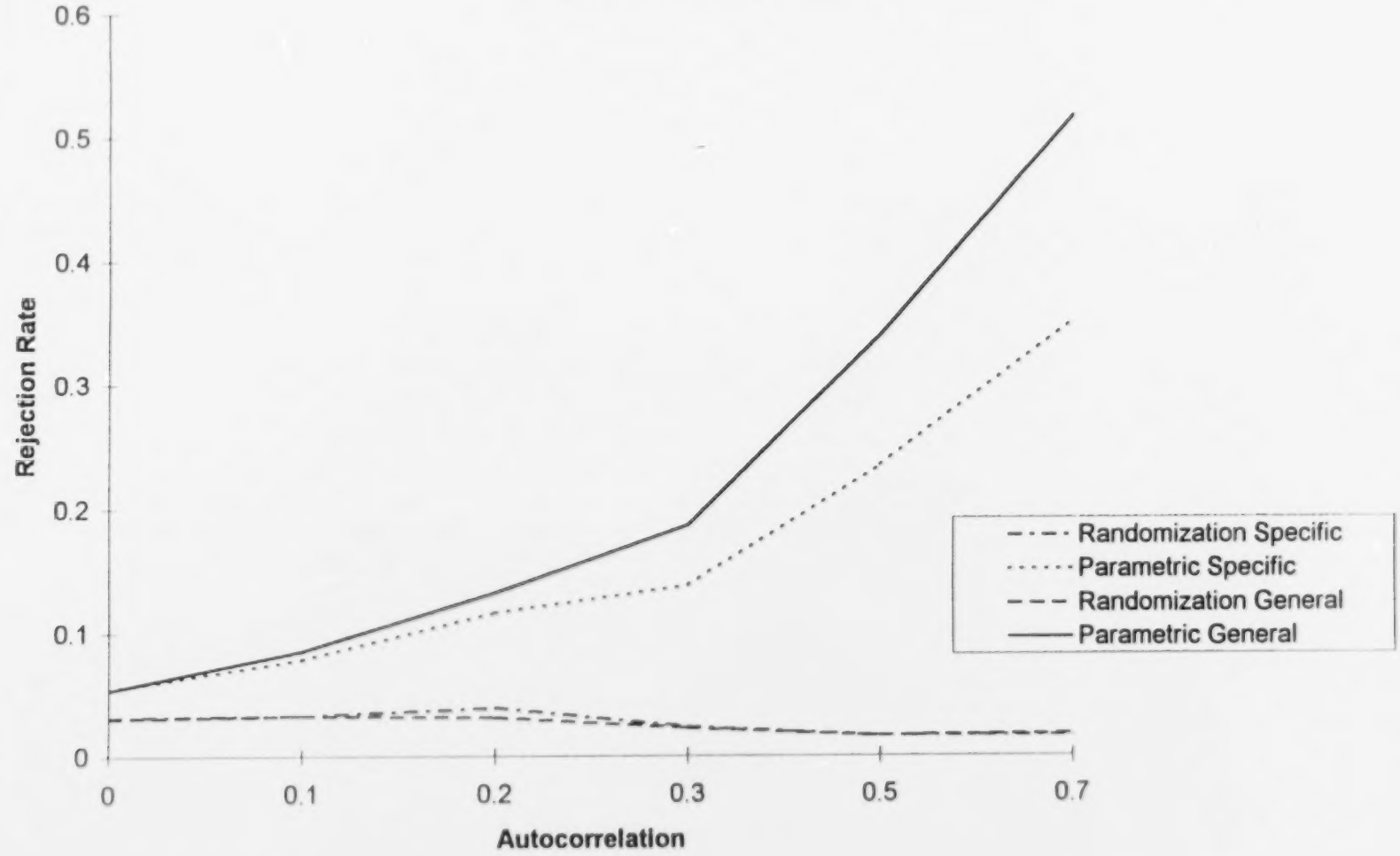


Figure 2. **Power Curves, Mean Shift Treatment Effects. Alpha=.05.**

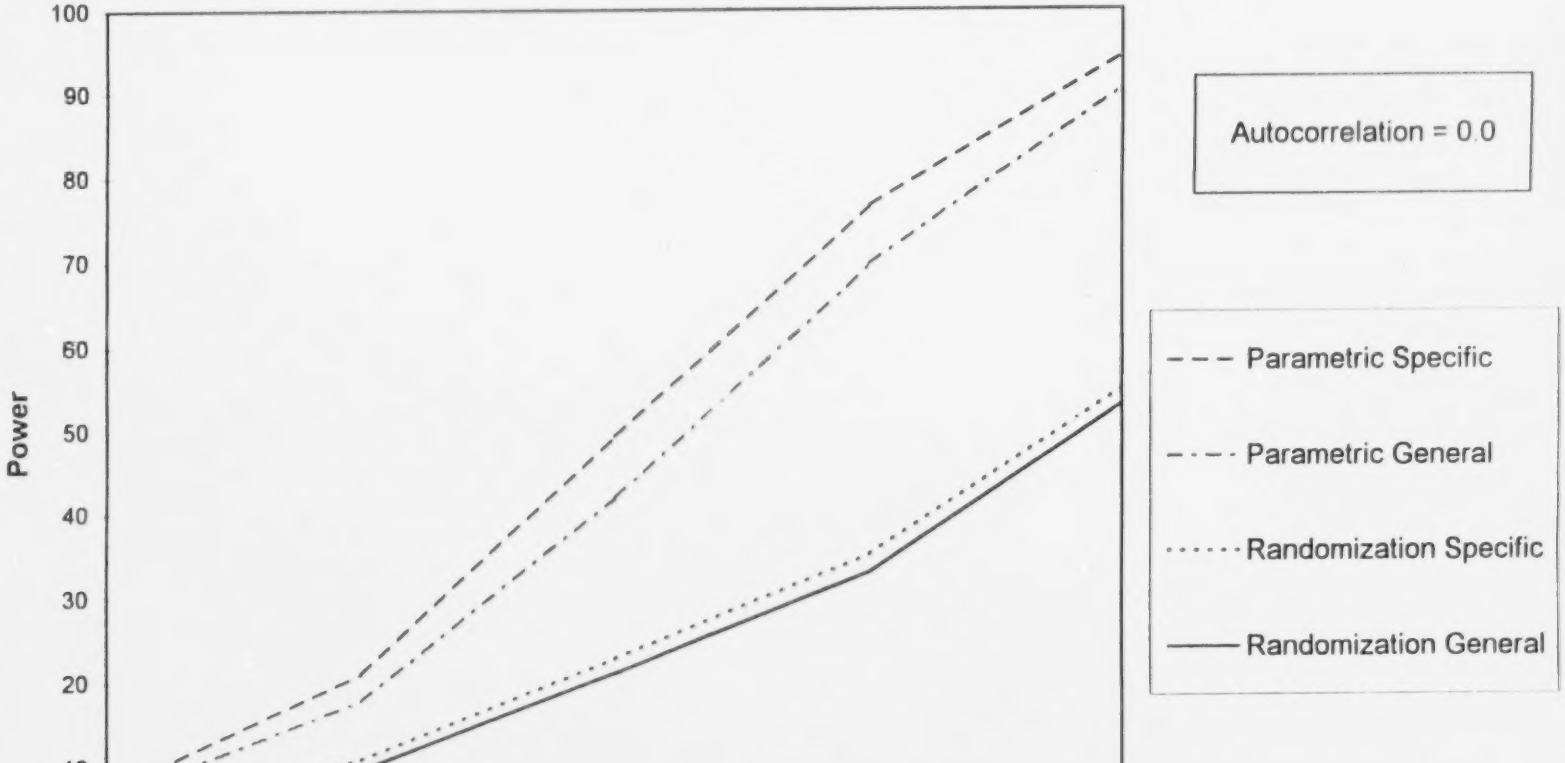


Figure 3. **Power Curves, Changing Slope Treatment Effects. Alpha=.05.**

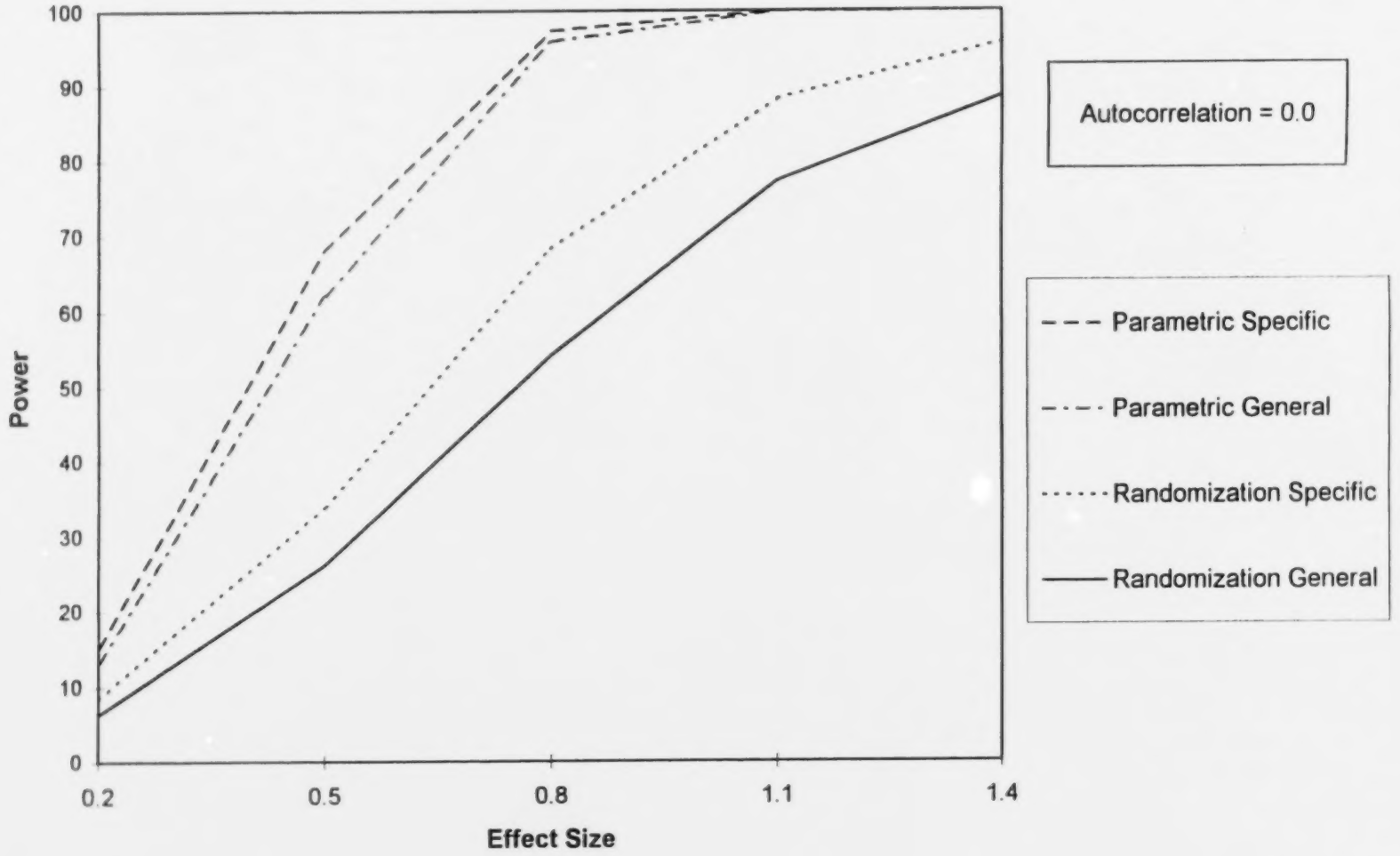
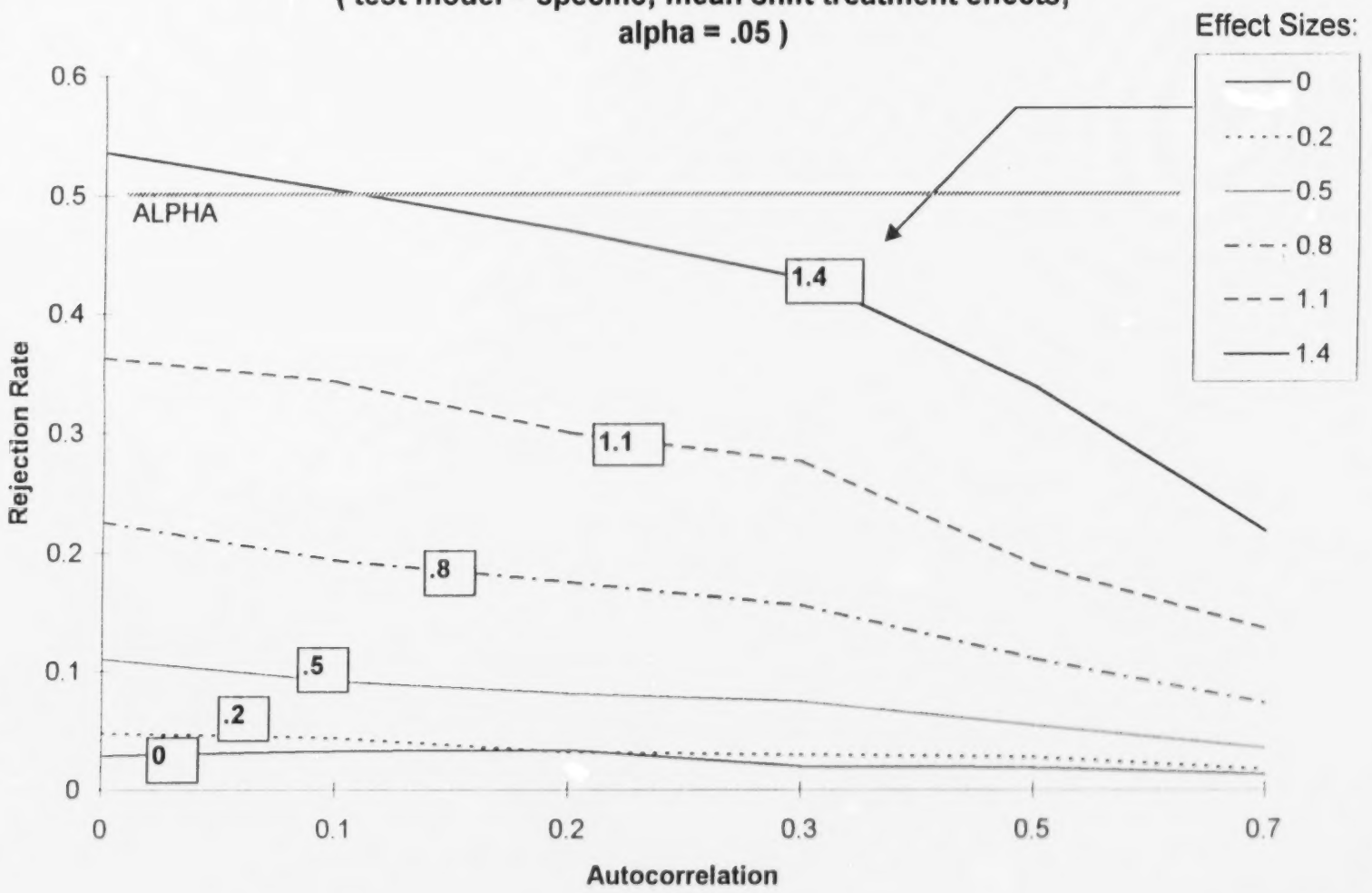


Figure 4.


Effect of Autocorrelation on Randomization Test Power
(test model = specific; mean shift treatment effects;
alpha = .05)
Effect Sizes:
0
0.2
0.5
0.8
1.1
1.4
ALPHA
Rejection Rate
Autocorrelation
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.5
0.7
1.4
1.1
.8
.5
.2
0